Stock Market Forecasting Using Machine Learning Algorithms

Shunrong Shen, Haomiao Jiang
Department of Electrical Engineering
Stanford University
{conank,hjiang36}@stanford.edu

Tongda Zhang
Department of Electrical Engineering
Stanford University
tdzhang@stanford.edu

***Abstract*—Prediction of stock market is a long-time attractive topic to researchers from different fields. In particular, numerous studies have been conducted to predict the movement of stock market using machine learning algorithms such as support vector machine (SVM) and reinforcement learning. In this project, we propose a new prediction algorithm that exploits the temporal correlation among global stock markets and various financial products to predict the next-day stock trend with the aid of SVM. Numerical results indicate a prediction accuracy of 74.4% in NASDAQ, 76% in S&P500 and 77.6% in DJIA. The same algorithm is also applied with different regression algorithms to trace the actual increment in the markets. Finally, a simple trading model is established to study the performance of the proposed prediction algorithm against other benchmarks.**

I. Introduction

Prediction of stock trend has long been an intriguing topic and is extensively studied by researchers from different fields. Machine learning, a well-established algorithm in a wide range of applications, has been extensively studied for its potentials in prediction of financial markets. Popular algorithms, including support vector machine (SVM) and reinforcement learning, have been reported to be quite effective in tracing the stock market and help maximizing the profit of stock option purchase while keep the risk low [1-2]. However, in many of these literatures, the features selected for the inputs to the machine learning algorithms are mostly derived from the data within the same market under concern. Such isolation leaves out important information carried by other entities and make the prediction result more vulnerable to local perturbations. Efforts have been done to break the boundaries by incorporating external information through fresh financial news or personal internet posts such as Twitter. These approaches, known as sentiment analysis, replies on the attitudes of several key figures or successful analysts in the markets to interpolate the minds of general investors. Despite its success in some occasions, sentiment analysis may fail when some of the people are biased, or positive opinions follow past good performance instead of suggesting promising future markets.

In this project, we propose the use of global stock data in associate with data of other financial products as the input features to machine learning algorithms such as SVM. In particular, we are interested in the correlation between the closing prices of the markets that stop trading right before or at the beginning of US markets. As the connections between worldwide economies are tightened by globalization, external perturbations to the financial markets are no longer domestic. It is to our belief that data of oversea stock and other financial markets, especially those having strong temporal correlation with the upcoming US trading day, should be useful to machine learning based predictor, and our speculation is verified by numerical results.

~~The rest of the report is organized as following. Section II presents our algorithm in details, including the fundamental principle of our algorithm, data collection and feature selection. Numerical results are shown in Section III followed by analysis and discussions. In Section IV, we established a simple trading model to demonstrate the capability of the proposed algorithm in increasing profit in NASDAQ. Section V summarizes the whole report.~~

II. Algorithms

A. Basic Principles

Globalization deepens the interaction between the financial markets around the world. Shock wave of US financial crisis hit the economy of almost every country and debt crisis originated in Greece brought down all major stock indices. Nowadays, no financial market is isolated. Economic data, political perturbation and any other oversea affairs could cause dramatic fluctuation in domestic markets. Therefore, in this project, we propose to use world major stock indices as input features for our machine learning based predictor. In particular, the oversea markets that closes right before or at the beginning of the US market trading should provide valuable information on the trend of coming US trading day, as their movements already account for possible market sentiment on latest economic news or response to progress in major world affairs.

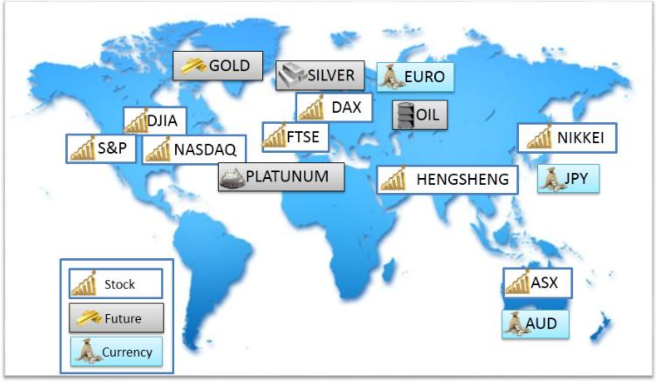

Fig. 1. World financial markets.

In addition to stock markets, commodity prices and foreign currency data are also listed as potential features, as different financial markets are interconnected. For instance, slowdown in US economy will definitely cause a drop in US stock market. But at the same time, USD and JPY will increase with respect to its peers as people seek for asset havens. Such interplay implies the underlying relationship between these financial products and the possibility of using one or some of them to predict the move of the other ones.

B. Data collection

The data set used in this project is collected from [3]. It contains 16 sources as listed in Table I and covers daily price from 04-Jan-2000 to 25- Oct -2012: Since the markets are closed on holidays which vary from country to country, we use NASDAQ as a basis for data alignment and missing data in other data sources is replaced by linear interpolation.

TABLE I. DATA SOURCE

Stock	NASDAQ, DJIA, S&P 500, Nikkei 225, Hang Seng index, FTSE100, DAX, ASX
Currency	EUR, AUD, JPY, USD
Commodity	Silver, Platinum, Oil, GoldC:\Dropbox\CS229 Project

C. Feature selection

In this project, we focus on the prediction of the trend of stock market (either increase or decrease). Therefore, the change of a feature over time is more important than the absolute value of each feature. We define $x_i(t)$, where $i \in \{1,2,..., 16\}$, to be feature i at time t. The feature matrix is given by

$$F = (X_1, X_2, ..., X_n)^T \tag{1}$$

where

$$X_t = (x_1(t),x_2(t),...,x_{16}(t)) \tag{2}$$

The new feature which is the difference between two daily prices can be calculated by

$$\nabla\delta x_i(t) = x_i(t) - x_i(t-\delta) \tag{3}$$

$$\begin{aligned}\nabla_\delta X(t) &= X(t) - X(t-\delta)\\ &= (\nabla_\delta x_1(t), \nabla_\delta x_2(t), \cdots \nabla_\delta x_{16}(t))^T \end{aligned} \tag{4}$$

$$\nabla_\delta F = (\nabla_\delta X(\delta+1), \nabla_\delta X(\delta+2), ..., \nabla_\delta X(n)) \tag{5}$$

Due to the difference in market value and basis of each market, the differential values calculated above can vary in a wide range. To make them comparable, the features are normalized as following:

$$\begin{aligned}\mathcal{N}(\nabla_\delta x_i(t)) &= \frac{x_i(t)-x_i(t-\delta)}{x_i(t-\delta)}\\ \mathcal{N}(\nabla_\delta X(t)) &= (\mathcal{N}(\nabla_\delta x_1(t)),\cdots,\mathcal{N}(\nabla_\delta x_{16}(t)))^T\\ \mathcal{N}(\nabla_\delta(\mathcal{F})) &= (\mathcal{N}(\nabla_\delta X(\delta+1)),\cdots,\mathcal{N}(\nabla_\delta X(n)))^T \end{aligned} \tag{6}$$

and the normalization can be implemented as:

$$normal(X(t)) = \frac{\mathcal{N}(\nabla_\delta X(t))}{|\mathcal{N}(\nabla_\delta X(t))|} \tag{7}$$

As discussed above, the performance of a stock market predictor heavily depends on the correlation between the data used for training and the current input for prediction. Intuitively, if the trend of stock price is always an extension to yesterday, the accuracy of prediction should be fairly high. To select input features with high temporal correlation, we calculated the autocorrelation and cross-correlation of different market trends (increase or decrease). The results shown in Figure 2 use NASDAQ as the base market.

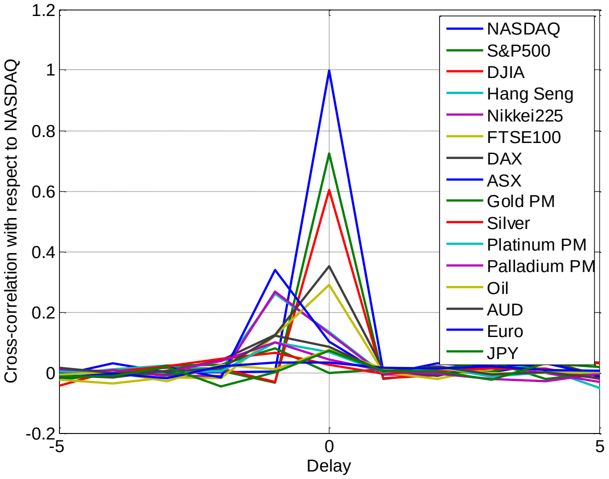

Fig. 2. Autocorrelation and cross-correlation of market trends using NASDAQ as a base.

It can be seen from the graph that the autocorrelation of NASDAQ daily trend is only non-zero at the origin based on which we may conclude that the trend of NASDAQ daily index is approximately a Markov process. Hence, past data of NASDAQ will not provide much insight to its future movement. The same conclusion could be made on many other data sources whose cross-correlation with NASDAQ is close to zero. Although the trend of DJIA and S&P500 have strong correlation with NASDAQ given the data are on same day, they are not available by the time they are needed for prediction. However, data sources such as DAX AUD and some other markets are promising features for our predictor built by machine learning algorithm as they have relative high correlation with NASDAQ at the origin and their data is available before or at the beginning of the US market trading time. This observation actually confirms our forecast principle, as discussed in previous sections, about the inter-connection between global markets and how the information reflected by their movements can be beneficial to the prediction of US stock markets.

In addition to the correlation of daily market movements, it is also worth investigating the correlation of market trends in longer term, which may also offer valuable information for predicting future price [4-5]. To study this, δ in Eq. (3) is varied from 1 day to 50 days, and part of the results are plotted in Figure 3. As can be seen from the graph, temporal correlations between the markets increase with the time window δ over which the trends of stock indices are calculated. One explanation for this phenomenon is that the calculation in Eq. (3) makes the time spans of the outputs overlap with each other and therefore increase the temporal correlations. Moreover, the operation also implicitly conduct an averaging on the data within the interval which effectively removes the noise and the underlying correlation between the markets become clearer.

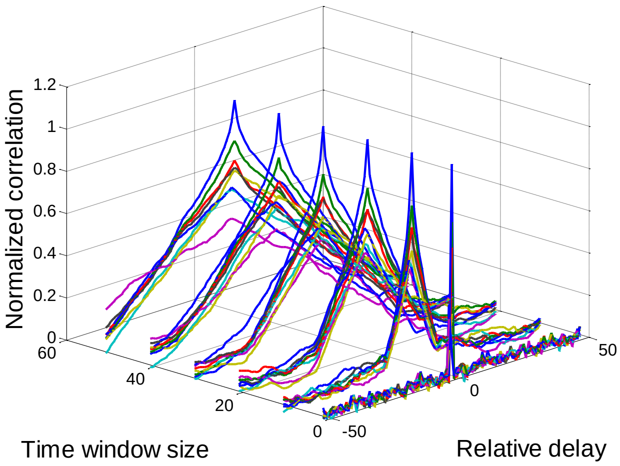

Fig. 3. Autocorrelation and cross-correlation of market trends with different time spans.

Given all possible features, we implemented forward feature selection algorithm to select the features that contribute most to the accuracy of prediction using different machine learning algorithms. Detail results are presented in the next section. As expected, a combination of daily market trend and long term movement provide the best result.

III. Experimental Results and Discussion

A. Trend Prediction

1) Single Feature Prediction

In section 2, we used cross-correlation to estimate the importance of each feature. To verify the information given by correlation analysis, we use individual feature to predict daily NASDAQ index trend. The prediction accuracy by each single feature is shown in Figure 4.

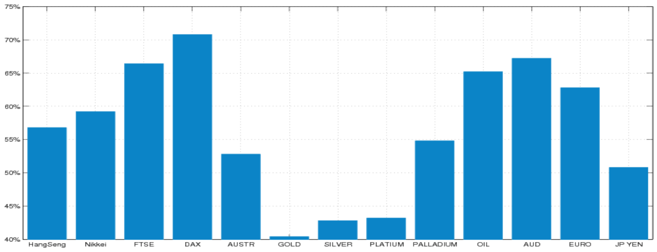

Fig. 4. Prediction accuracy by single feature

From the figure, we can see that DAX yields the best results, 70.8% accuracy of prediction. Prediction accuracies of Australian dollar, FTSE and oil price are also relatively high, reaching 67.2%, 66.4% and 65.2% respectively.

The result of this experiment supports the analysis of cross-correlation. Hence, we are convinced that index value of other stock markets and commodity prices can provide useful information in the prediction process.

2) Long Term Prediction

Besides daily movement, sometimes we also care about prediction results for longer terms. Here, we define our problem as predicting the sign of difference between tomorrow's index value with respect to that of certain days ago. We use SVM as a training model and the prediction accuracies under different time spans are shown in Figure 5.

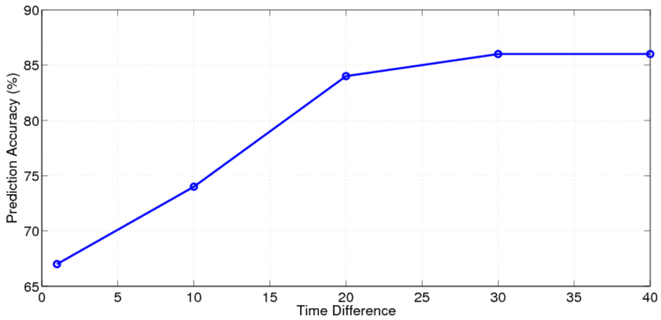

Fig. 5. Accuracy of Long Term Prediction

From the figure, we can see that prediction accuracy increases when time span becomes longer. This is because that the longer the period is, the more information we have and the higher resistance of our prediction to noise. At last, we can reach 85% accuracy when time spam gets longer than 30 days.

Actually, we can re-paraphrase this problem as estimation of $\Pr\{V_{t+1} - V_t > C_t\}$, where $C_t = -(V_{t-t_s} - V_t)$. This corresponds to the after-mentioned regression problem of daily stock market movements.

3) Multi-Feature Prediction

Using the features described in section 2, we compare the prediction accuracies of SVM algorithm and MART (a decision tree based boosting algorithm). The prediction results are shown in Table 2.

TABLE II. One-day Prediction Accuracy

	Top 4 Features	All Features
SVM	74.4%	63.1%
MART	70.3%	73.9%

From Table 1, we can see that accuracies from SVM and MART learner can reach as high as 74%. This daily trend prediction accuracy is higher than most of models and the values reported on financial analysis websites.

In addition, we note that SVM algorithm is very sensitive to the size of training data. When the size of training set is not large enough, hyper-plane found by SVM algorithm might not be able to split the data properly. Thus, feature selection is essential when using SVM. In contrast, Multiple Additive Regression Trees (MART) algorithm requires less training data and prefers high dimensional feature set.

To test the generality of our model, we used the same algorithm to predict the other two US stock markets. The results are shown in Table 3 below.

TABLE III. PREDICTION ACCURACY ON ALL US STOCK MARKETS

	NASDAQ	DJIA	S&P 500
Accuracy	74.4%	77.6%	76%

As can be seen, all entries in the table are high. This shows that our model can be applied to all US stock markets. Actually, the idea of using time delay between different stock markets can be also utilized to predict other indexes.

~~*B. Regression*~~

~~Compared to stock trend, the exact increment in stock index may provide more information for investment strategy. This means the classification problem now evolves to a regression problem. To judge the performance of our model, we use square root of mean square error (RMSE) as criteria, which is defined as~~

$$\mathrm{RMSE} = \left(\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - y_i)^2\right)^{\frac{1}{2}}$$
~~(6)~~

~~We use linear regression, generalized linear model (GLM) and SVM algorithm to predict exact value of daily NASDAQ movement. The RMSE values for different algorithms are reported in Table 4 below.~~

~~TABLE IV. STOCK INDEX REGRESSION ACCURACY~~

	~~Baseline~~	~~SVM~~	~~Linear~~	~~GLM~~
~~RMSE~~	~~40.4~~	~~21.6~~	~~24.8~~	~~28.7~~

~~The baseline predictor in the table is formed by a zero-order hold filter. From the table, we can see that SVM gives the most accurate prediction. The RMSE given by SVM is 21.6, only half of the average fluctuation, 47.66.~~

C. Multiclass Classification

In previous part, we explored various methods to improve prediction accuracy and minimize square root mean square error. These efforts can be directly used to maximize the trading profit. However, besides maximizing profit, another aspect of our task is to minimize trading risk. In this part, we will use the SVM regression model and start from the basic intuition in SVM algorithm.

In SVM, the further distance between the point and hyper-plane is, more confident we are for the prediction we made, whereas, our prediction cannot be very accurate when the point is close to hyper-plane. To minimize the trading risk, we can pick out these risky points and ignore their prediction labels. Thus, we need to classify the original data into at least three classes, negative, neutral and positive. This is the intuition that leads to the prototype of our multiclass classification model.

To build up the multiclass classifier, we firstly need to define the width of the central region. To evaluate how well our classifier is, we introduced the concepts of precision and recall, which are defined as

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}} \qquad (7)$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp+fn}} \qquad (8)$$

In the equations above, tp, fp and fn stand for true positive, false positive and false negative respectively. The precision and recall values against different central region widths are plotted in Fig. 6. In the figure, Recall for positive class reflects the proportion of predicted rising days among all rising days while precision indicates the hit rate among rising predictions. Thus, recall directly impacts the frequency of trading and precision impacts the profit / loss at each time. Taking the product of the two, we calculate the F1 score which is defined as

$$\text{F1} = 2\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \qquad (9)$$

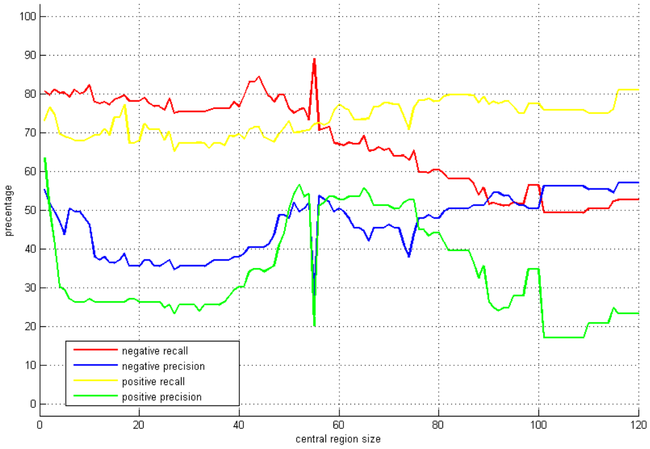

Fig. 6. Precision and recall vs. central window widths for positive and negtive class

The value of F1 scores for positive and negative classes are shown in Fig. 7. As can be seen from the graph, F1 scores for both class are relatively high around 0 and 50. Therefore, we shall choose window size around 0 when trading fees / tax are small enough to be ignored. Otherwise, we should choose 50 as the optimal window size.

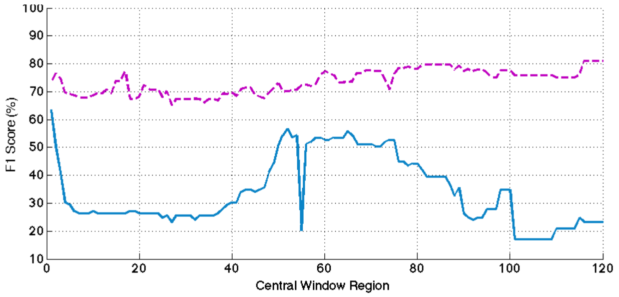

Fig. 7. F1 score vs. central window size for positive and negative class

IV. Trading Model

In this section, we will build up a trading model based on the predictor we find in section 2 and section 3. We compare the simulation results of our model against two selected benchmarks.

A. Basic Settings

We randomly select 5 different time slots for simulation, 50 days inside each time slots. The initial capital at the beginning of the each time slot is 10,000 dollars. At the end of the trading period, all possessed stocks are forced to be cashed. Furthermore, for simplicity, we suppose that there is no Stamp Duty or any kind of tax or fees during the process and short sale is not allowed in our simulation.

B. Model Specification

In our simulation, we use two benchmark models and one model using our predictor. Here, we will describe the three models in detail.

1) Benchmark model 1

In this simple model, we suppose we use all the money to buy stocks on the first day and sell the stocks at the end. Thus, the profit is depend on the trend during this testing period, that is

$$\text{Profit} = \frac{10{,}000}{V_0}(V_t - V_0) \qquad (10)$$

2) Benchmark model 2

In this model, we assume that closing stock index of tomorrow is higher than today if today's index is higher than that of yesterday. Whenever the prediction is rising, we buy at most S shares of stocks. Otherwise, we sell all stocks we have.

This model performs well when stock markets go smoothly. But it losses a lot when the markets fluctuate or shocks frequently.

3) Proposed Trading Model

We use the prediction results from our SVM learner. The trading principal is the same as Benchmark model 2. That is, we buy stock when prediction is positive and cash all stocks we have when prediction is negative.

C. Simulation Results

The profit of the three models during the five period is shown in Table 5 below. From the table, we can see that during most of periods, our proposed model wins the most profits. On average, our model gains 814.6 dollars as profits for every 50 days. That is 8% return rate in 50 days. Therefore, we can reach annual interests at about 30%.

Besides high profit, our model also has the advantage of low risk. Our model seldom loses in trading period while benchmark model 1 and benchmark model 2 loses in period 3 and 5. Actually, in most cases, our model can get at least 5% profits in the 50 day long trading period.

Although we can reach high profit and low risk in our simulation, we shall still remember that we did not take tax and fees into consideration. We also assume that our operation will not have direct effects on index value.

TABLE V. Trading Results of Different Trading Strategies

	Benchmark 1	Benchmark 2	Our Model
Period 1	98.16	-638.45	1539.63
Period 2	1259.66	1281.82	685.27
Period 3	-1103.13	-524.43	921.81
Period 4	654.58	0.37	894.13
Period 5	-249.05	-171.08	32.18
Average	132.04	-10.35	814.60

V. Conclusion

In the project, we proposed the use of data collected from different global financial markets with machine learning algorithms to predict the stock index movements. Our findings can be summarized into three aspects:

1. Correlation analysis indicates strong interconnection between the US stock index and global markets that close right before or at the very beginning of US trading time.
2. Various machine learning based models are proposed for predicting daily trend of US stocks. Numerical results suggests high accuracy
3. A practical trading model is built upon our well trained predictor. The model generates higher profit compared to selected benchmarks.

There are a number of further directions can be investigated starting from this project. The first one is to explore other creative and effective methods that might yield even better performance on stock market forecasting. Second, models can be modified to take care of the tax and fees in the trading process. Finally, we can investigate the mechanism of short sale and maximize our profit even when the market is bullish.

Reference

[1] W. Huang et al., "Forecasting stock market movement direction with support vector machine," Computers & Operations Research, 32, pp. 2513–2522005, 2005

[2] J. Moody, et al., "Learning to trade via direct reinforcement," IEEE Transactions on Neural Networks, vol. 12, no. 4, Jul. 2001.

[3] www.wikiposit.org

[4] S. Zemke, "On developing a financial prediction system: Pitfall and possibilities," Proceedings of DMLL-2002 Workshop, ICML, Sydney, Australia, 2002.

[5] Vatsal H. Shah, "Machine learning techniques for stock prediction," www.vatsals.com.